SMARTER: A Data-efficient Framework to Improve Toxicity Detection with Explanation via Self-augmenting Large Language Models

Anonymous ACL submission

Abstract

WARNING: *This paper contains examples of offensive materials.* To address the proliferation of toxic content on social media, we introduce *SMARTER*, a data-efficient 2-stage framework for explainable content moderation using Large Language Models (LLMs). In Stage 1, we leverage LLMs' own outputs to generate synthetic explanations for both correct and incorrect labels, enabling alignment via preference optimization with minimal human supervision. In Stage 2, we refine explanation quality through cross-model training, allowing weaker models to align stylistically and semantically with stronger ones. Experiments on 3 benchmark tasks (*HateXplain, Latent Hate, and Implicit Hate*) demonstrate that SMARTER enables LLMs to achieve up to 13.5% macro-F1 improvement over standard few-shot baselines with a fraction of the full training data. Our framework offers a scalable strategy for low-resource settings by harnessing LLMs' self-improving capabilities for both classification and explanation.

1 Introduction

In recent years, social media platforms have enabled millions of users to virtually interact at a global scale. While a vital channel to disseminate information and knowledge, these platforms also facilitate the spread of harmful materials (Hwang and Kim, 2015; Cinelli et al., 2021; Nguyen et al., 2024). Arguably the most concerning among them is toxic content, which could induce serious negative psychological impact on the audience (Castaño-Pulgarín et al., 2021; Windisch et al., 2022; Tontodimamma et al., 2021). As a result, content moderation is crucial to detect and mitigate its harm.

Toxic content encompasses a wide spectrum of terminologies whose definitions vary by platform: hate speech, cyberbully, sexist, racist etc. (Gelber, 2021; Anjum and Katarya, 2024). Typically, human moderators must manually adjudicate the target content, an inefficient process taxing to the moderators themselves (Baker et al., 2020; Spence et al., 2023). Recently, social media platforms can significantly reduce this overhead by integrating machine learning models to automate the detection pipeline (d'Sa et al., 2020; Sumanth et al., 2022).

While capable of achieving impressive classification performance, these increasingly sophisticated models require ample resources to train, especially when the target label space covers nuanced concepts (Fortuna et al., 2021; Sap et al., 2021; Zhao et al., 2021). Developing a robust dataset for training typically demands significant effort in curation and human annotation, often involving iterative updates to adapt the model to evolving linguistic trends (Toraman et al., 2022; Bespalov et al., 2023; Rad et al., 2025). In addition, this pipeline is limited to decision-only pipeline. Failure to provide users with human-understandable signals renders the adjudication process opaque, inviting further issues on transparency and trustworthiness.

In this paper, we address the pressing need for intuitive explanations as a cornerstone of content moderation by leveraging Large Language Models (LLMs), which have shown remarkable capabilities in reasoning-related tasks (Wei et al., 2021; Yao et al., 2024). Specifically, our contributions are:

- ◇ We introduce **SMARTER**[1] (Self-augMentAtion Regimen Towards Efficient Content ModeRation), a 2-stage framework where (1) each LLM self-augments on few-shot data using preference optimization with synthetic explanations, (2) cross-model refinement to further improvement by training one LLM with responses generated by another on held-out sets.
- ◇ Experiments conducted on 3 benchmark datasets (*HateXplain, Latent Hate, Implicit Hate*) with Llama-3.1-8B-Instruct and COT-

[1] Our repository can be found at REDACTED

T5-XL demonstrate up to 13.5% increase in macro-F1 scores in few-shot settings with high-quality explanations, outperforming a range of commercial models.

SMARTER thus provides a *scalable and practical approach for low-resource NLP practitioners* to achieve strong performance and explainability.

2 Related Works

Robust Toxicity Detection Existing research (Caselli et al., 2021; Sarkar et al., 2021; Toraman et al., 2022) has proposed strategies to enhance cross-domain robustness in toxicity detection, such as combining multiple datasets to improve generalization (Antypas and Camacho-Collados, 2023) and incorporating label definitions to support few-shot adaptation (Nghiem et al., 2024). Building on these advances, recent studies leverage LLMs for toxicity detection, demonstrating strong zero- and few-shot performance (Masud et al., 2024; Jahan et al., 2024; Roy et al., 2023; Kumarage et al., 2024). LLMs can also generate textual explanations, enhancing interpretability for moderators and users (Di Bonaventura et al., 2024; Yang et al., 2023). We extend this line of work by training open-source LLMs to jointly perform classification and explanation to promote transparency in moderation workflows.

Aligning LLMs to Human Preferences In addition to extensive pretraining, modern LLMs also undergo post-training that both unlocks additional capabilities (Longpre et al., 2023) and aligns the outputs towards human preferences (Ouyang et al., 2022). Recent Reinforcement Learning (RL)-based techniques like GRPO (Shao et al., 2024; Zweiger et al., 2025) enable promising LLM self-alignment at expensive computational cost. Another body of work showcases LLMs to self-improve via their own synthetic data (Tang et al., 2023; Li et al., 2023), such as Self-Refine (Madaan et al., 2023) and Self-Instruct (Wang et al., 2022). Building upon these ideas, our framework presents a simple yet principled approach that allows LLMs to iteratively improve with minimal training cost while still amenable to human supervision.

3 Overview of Framework

As shown in Algorithm 1, our framework consists of two stages. First, we focus on optimizing the LLMs' classification performance through alignment tuning using self-augmenting synthetic data generation. In the second stage, we refine the quality of the generated explanations by cross-model refinement, leveraging the synergy between different LLMs. Section 4 and 5 present the experimental setup and findings for Stage 1. Section 6 focuses on cross-model improvement in Stage 2. Section 7 analyzes the consistency of explanations with respect to the definitions and labels.

Algorithm 1 High-Level Overview of SMARTER

Input: LLMs $\mathcal{L}$, performance metric $\mathcal{M}$
Datasets $\mathcal{D}_{\text{train}}$ (with explanation), $\mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$,
Optional: Initial SFT pretraining on relevant data
Stage 1: Individual LLM Self-Augmentation
for $K \in \{K_{\text{start}}, ..., K_{\text{end}}\}$ **do**
  Sample K-shot subset $\mathcal{D}^{(K)}$
  for all $\text{LLM}_i \in \mathcal{L}$ **do**
    SFT on $\mathcal{D}^{(K)}$
    Collect *chosen* responses for gold labels
    Collect *rejected* responses for incorrect labels
    Align via preference optimization using synthetic preference data; evaluate on $\mathcal{D}_{\text{val}}$
  end for
end for
Stage 2: Cross-Model Refinement
for $\text{LLM}_a \neq \text{LLM}_b \in \mathcal{L}$ **do**
  Select $K_{check} \in \{K_{\text{start}}, ..., K_{\text{end}}\} \rightarrow \mathcal{D}^{(K_{check})}$
  SFT LLM_a on $\mathcal{D}^{(K_{check})}$
  Self-augment LLM_a with responses from LLM_b on $\mathcal{D}^{(K'_{check})} \neq \mathcal{D}^{(K_{check})}$
end for
return Model with best performance $\mathcal{M}$ on $\mathcal{D}_{val/test}$

4 Preliminary Setup

In this section, we provide details about the empirical setup on data, model, and a description of classification technique via prompting.

4.1 Data

We select 2 benchmark datasets that cover distinct yet still relevant concepts to toxic content moderation and derive from them 3 classification tasks.

- **HateXplain**, introduced by Mathew et al. (2021). Consisting of posts collected from 2 social media sites Gab and Twitter (now X), this dataset contains 20,148 samples annotated by crowdworkers of the Amazon Mechanical Turks (AMT) platform on 3 labels *Hate, Offensive* and *Normal*.
- **Latent Hate**, curated by ElSherief et al. (2021), focuses on covert hate speech collected from Twitter. With 19,112 samples also annotated by AMT workers, this dataset's primary label space contains 3 high-level categories: *Not Hate, Implicit Hate, Explicit Hate*.

The authors also assigned the subset of 4,153 tweets of the *Implicit Hate* category a label among 6 secondary fine-grained categories: *White Grievance, Incitement to Violence, Inferiority Language, Irony, Stereotypes and Misinformation, Threatening and Intimidation*.

Definitions of the associated labels are also included, allowing us to integrate these contexts into the explanation in downstream applications in a fashion similar to Yang et al. (2023) and Nghiem and Daumé (2024). We refer to the three classification tasks by their label spaces as *HateXplain*, *Latent Hate*, and *Implicit Hate*. Table 4 summarizes their components and data splits, with preprocessing details in Appendix B.

4.2 Models

Extending our focus on accessibility, we employ 2 open-source LLMs that have exhibited strong performance in similar classification tasks:

- ◇ **COT-T5-XL**: A variant of the base encoder-decoder Flan-T5-XL (Longpre et al., 2023) architecture, this model consists of 3 billion parameters and is further trained on the COT corpus, a collection of 1.8 million samples with chain-of-thought style explanations, which enhances its capabilities on a variety of reasoning tasks (Kim et al., 2023).
- ◇ **Llama-3.1-8B-Instruct**: A decoder-only LLM released by Meta (Dubey et al., 2024). Consisting of 8 billion parameters, this model is further instruction-tuned to aligned with human preferences in addition to being trained on extensive linguistic corpora.

For brevity, we refer to the models as *T5* and *Llama* respectively for the remainder of this paper.

4.3 Classification via Prompting

We apply this Chain-of-Thought (COT) paradigm (Wei et al., 2022) in our classification tasks to obtain the explanation from the LLMs necessary for content moderation. Inspired by Nghiem and Daumé (2024) and Yang et al. (2023), we use the prompt in Figure 6, which incorporates the target post, the set of labels corresponding to the task and their associated definitions as context for the LLM to provide its explanation and predicted category.

4.4 Auxiliary Pretraining

Nghiem and Daumé (2024) found that supervised finetuning (SFT) off-the-shelf LLMs on *HateCOT* – a general compilation of toxic posts, labels, definitions and synthetic explanations – can boost their performance in downstream hate speech detection. To maximize expected gains with minimal in-domain data, we adopt this strategy and perform SFT on the 2 chosen LLMs using *HateCOT* to prime them for our classification tasks. Finetuning is performed by training LoRA adapters (Hu et al., 2021), a parameter-efficient approach implemented by the HuggingFace library (see Appendix F) as the basis for downstream experiments.

5 Stage 1: Individual Model Training

5.1 K-shot In-domain Finetuning

We experiment with different configurations of limited in-domain data. For each task, we select uniformly at random K-shot $\in \{16, 32, 64, 128, 256\}$ samples from the training set . We utilize the seed explanations by Nghiem and Daumé (2024) to further SFT the existing *HateCOT*-pretrained *T5* and *Llama* models in a fashion similar to the auxiliary pretraining phase above.

Self-augmenting via Alignment Tuning We optimize the LLMs' performance while not having access to more training data by exploiting the models' sycophantic tendency. More concretely, we augment the training sample size by prompting the LLMs to generate explanations *conditioned on the incorrect labels* (as illustrated in Figure 4). This process supplies us with synthetic post-explanation pairs that are by design *dispreferred* compared to the true *preferred* sample pairs, enabling the use of the following alignment techniques to further finetune the models.

- ◇ **Direct Preference Optimization (DPO)**: An offline RL-based technique that optimizes the policy via an implicit reward model using preference data (Rafailov et al., 2024). DPO requires human preference data where a pair of chosen vs. rejected responses is presented for each prompt.
- ◇ **Kahneman-Tversky Optimization (KTO)**: An alternative to DPO, KTO leverages prospect theory to construct human-aware losses (HALOs) to optimize the policy (Ethayarajh et al., 2024). KTO only requires listwise human preference data: a data point consists of a prompt, a response and a binary flag that indicates the acceptability of the response.

Appendix F.3 offers more technical details on these techniques. Our self-augmenting method-

ology naturally procures training datasets in the required format for KTO by simply designating the explanation conditioned on the gold label as positive, and otherwise negative. For DPO, we match the explanations conditioned on the correct-incorrect pair sequentially to satisfy the pairwise chosen-rejected format. As an example, 100 posts of *HateXplain* (whose label space is of size 3), should produce $100 * 3 = 300$ training samples for KTO, and $100 * (3 - 1) = 200$ for DPO.

5.1.1 Experimental Pipeline

The experiments follow a two-stage process. First, we finetune *T5* and *Llama* under extreme data scarcity by setting $K \in \{16, 32, 64, 128\}$, followed by self-augmentation via DPO to compare classification performance. Then, we experiment with significantly more training data with $K = 256$ and evaluate KTO alignment against DPO [2]. For comparison, we also finetune both LLMs with the entire available training data to obtain classification label *without explanation*, denoted *Full* models. In addition, we also finetune ModernBERT (large) (Warner et al., 2024), a decoder-only model on the full training data and obtain the classification outputs using the same test sets.

In addition, we examine 2 partial sampling strategies to circumvent the proportionate growth of self-augmenting synthetic data to label space when K is large . In the first, we construct augmented data by selecting 128 and 192 shots of samples uniformly at random from the Baseline's 256-shot pool. In the second, instead of K-shot sampling, we select $K' * |S| * (|S| - 1)$ samples from all possible post-label pairs of the 256-shot Baseline pool (incorrect labels included) to create augmented data, where $K' \in \{128, 192\}$ and $|S|$ is the label space's size. While both approaches should yield identical size of training data for the same K', the first method, denoted *DPO-K*, preserves post-label parity whereas the second method, denoted *DPO-N*, yields higher diversity of post content.

Experiments are implemented with the HuggingFace library. We sample $K = 50$ shot on the held out Validation extraction of each dataset to tune hyperparamters, and evaluate on samples from the Test set as shown in Table 4. Technical details are included in Appendix F.

[2] Implementation for KTO trainer for *T5* is not available at the time of writing, so we are only able to perform this variant for *Llama*

5.2 Classification Results

5.2.1 Results for $K <= 128$

We report macro-F1–the harmonic mean between Precision and Recall–as the evaluation metric (Christen et al., 2023). Figure 1 displays the bar plots of the LLM's Baseline and DPO-enhanced performance on the test sets.

Self-augmentation with DPO generally boosts classification performance over its Baseline counterparts, with the sole reversal for *T5* on *Latent Hate* at $K = 16$ (2.1% reduction) along with static scores for *T5* on *HateXPlain* at $K = 32$, and for *Llama* in *Latent Hate* at $K = 64$. The infrequency of these observations and the negligible decrease highlight our method's consistent enhancement. In fact, increase in F1 scores go up to 14.5% for *Llama* (*HateXplain,* $K = 128$) and 25% for *T5* (*Implicit Hate*, $K = 64$).

Notably, DPO self-augmentation improves performance even at $K \leq 64$, showcasing our method's advantage in low-resource settings. *Llama* Baseline models appear to consistently outperform *T5* across all tasks. However, with DPO self-augmentation, *T5* is able to narrow the margin on *Latent Hate* and *Implicit Hate*. In contrast, *Llama* still dominates on *HateXplain*. These discrepancies suggest an interaction between dataset and model choice, and practitioners may benefit from contrasting different options with our self-augmenting techniques for best results.

5.2.2 Results for $K = 256$

In Figure 1, we observe the macro F1 scores for models $K = 256$ on the subfigures on the right. DPO-augmentation continues to enhance Baseline models which already benefits from more training data relative to lower K values, with improvement in scores reaching up to 13.2% for *LLama* and 14% for *T5*. In contrast, KTO-augmentation appears ineffectual on *HateXplain* and even hinders performance on *Latent Hate* and *Implicit Hate*, showing notable reduction from Baseline scores, respectively. With this, we hypothesize that DPO's contrastive nature reinforces signals that allow the LLM to learn from its own mistakes, thereby sharpening its reasoning ability for this type of discriminatory task. On the other hand, KTO's reliance on implicit signals from listwise data is not sufficient to improve discernment of classes when the semantic distinction between labels are nuanced for *Latent Hate* and *Implicit Hate*.

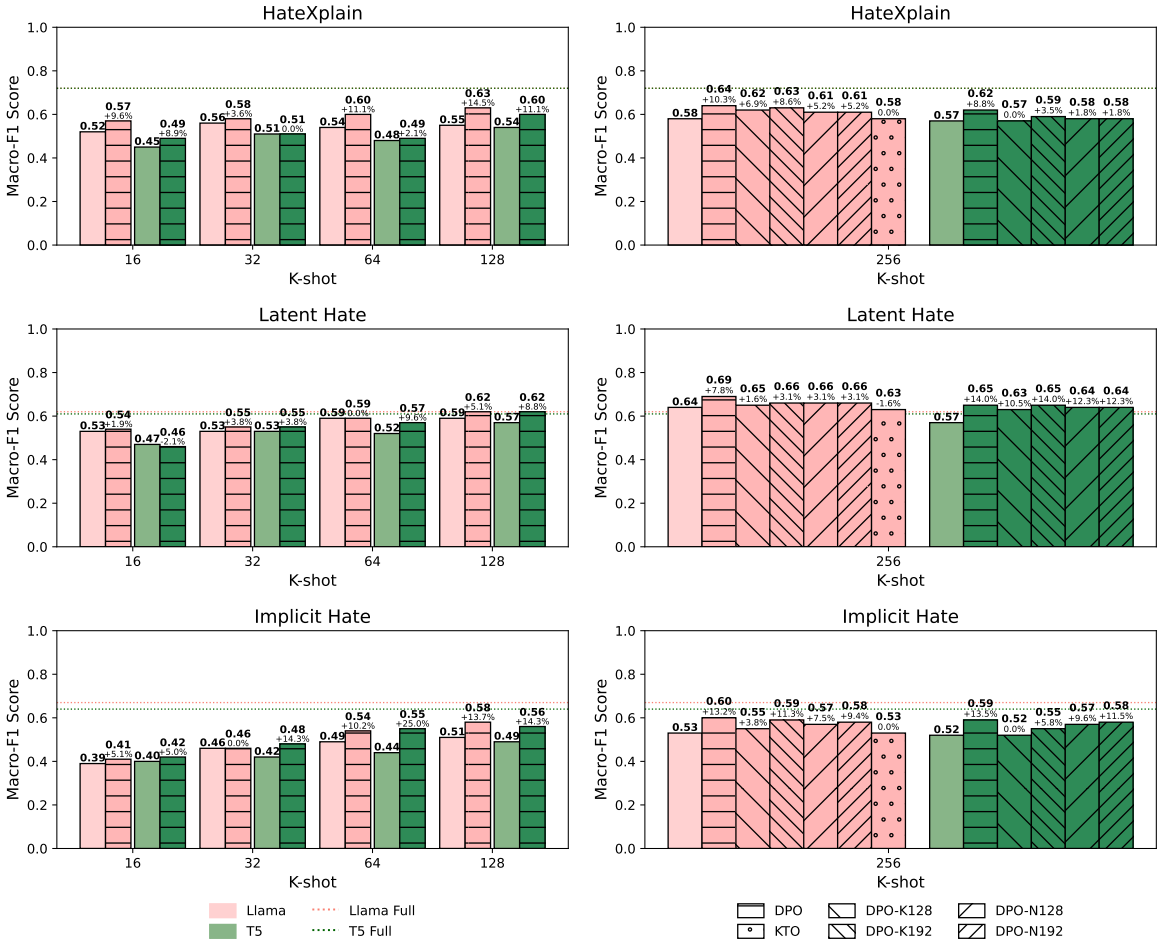


Figure 1: Bar plots for K-shot classification experiments on 3 datasets using *Llama* and *T5* models. Macro F1 scores and percentage change over Baseline are displayed on top. Results for Baselines and DPO-augmented variants for $K \in \{16, 32, 64, 128\}$ are displayed on the left subfigures. Results for $K = 256$ of Baseline, KTO, DPO-augmented and its other variants on various sub-sampling strategies (section 5.1.1) are shown on the right. Horizontal lines show the F1 scores for *Full* models that use all training data.

Results also show that DPO-augmented sub-sampling strategies still demonstrate generally better F1 scores than Baselines, further reinforcing the benefit of DPO-augmentation. There appears no clear winner between *DPO-K* and *DPO-N* among all 3 datasets. Augmenting with larger K' values (approaching 256) consistently yields scores closer to the maximal F1, indicating its role as a hyper-parameter that practitioners need to decide for the trade-off between computational cost and benefit.

Comparison against Alternative Baselines We further evaluate zero-shot classification on three *commercial* OpenAI LLMs and *Qwen-32B*, alongside off-the-shelf *T5* and *Llama* variants, using the prompt in Figure 5. As shown in Table 1 and Table 5, our DPO-augmented models at $K = 256$ consistently outperform both their baselines and larger LLMs, demonstrating SMARTER's accessibility advantage—achieving controllable, high-quality performance with lightweight, low-cost models that do not rely on proprietary platforms.

Following the same setup as *ModernBERT*, we also SFT-finetune base *Llama* and *T5* for classification *without* explanations. At $K = 256$, our DPO-augmented variants retain at least 86% of the *Full* model's F1 on *HateXplain* and surpass it on *Latent Hate*, underscoring the efficiency and scalability of our low-resource alignment framework.

6 Stage 2: Cross-model Refinement

In this section, we focus on investigating methods to identify and efficiently improve the LLMs' quality of explanations with respect to certain desired directions. We first enlist human annotators to evaluate the explanations by *T5* and *Llama*, then experiment with further training one model with the other's outputs to simultaneously improve performance and quality of explanations.

Dataset	Model	F1	F1%	Data%
HateXplain	Llama_DPO	0.64	89%	6%
	Llama_Full	**0.72**	100%	100%
	T5_DPO	0.62	86%	6%
	T5_Full	**0.72**	100%	100%
	ModernBERT	0.70	97%	100%
	GPT-4o-mini	0.50	69%	–
	GPT-4.1	0.55	76%	–
	GPT-5-chat	0.56	78%	–
	Qwen-32B	0.54	75%	–
Latent Hate	Llama_DPO	**0.69**	100%	7%
	Llama_Full	0.62	90%	100%
	T5_DPO	0.65	94%	7%
	T5_Full	0.61	88%	100%
	ModernBERT	0.61	88%	100%
	GPT-4o-mini	0.54	78%	–
	GPT-4.1	0.60	87%	–
	GPT-5-chat	0.51	74%	–
	Qwen-32B	0.47	68%	–
Implicit Hate	Llama_DPO	0.60	90%	57%
	Llama_Full	**0.67**	100%	100%
	T5_DPO	0.59	88%	57%
	T5_Full	0.64	96%	100%
	ModernBERT	0.64	96%	100%
	GPT-4o-mini	0.42	63%	–
	GPT-4.1	0.57	85%	–
	GPT-5-chat	0.58	87%	–
	Qwen-32B	0.49	73%	–

Table 1: Comparison of performance and data usage across models (including additional GPT baselines). We report Macro F1 (**F1**), relative score (**F1%**), and training data fraction (**Data%**). Models with the DPO suffix are DPO-augmented with K=256 shots, while *Full* uses the entire training set. Within each dataset, **F1** is bolded for the best model, and **F1%** is normalized to that best model (100%). **Data%** is measured relative to the corresponding *Full* baseline (100%). ModernBERT is shown for reference.

6.1 Human Evaluation

Due to resource constraints, we only perform this evaluation pipeline for *HateXplain*. We opt to collect outputs from the SFT+DPO models at $K = 128$ to preserve the rest of the training data for further refinement experiments. To make a fair comparison on the outputs to agree on the predicted label, we use the prompt in Figure 7 to obtain explanations conditioned on the correct gold label. This technique allows us to obtain 342 samples with consistent labels, or 89% of the possible $128 * 3 = 384$ possible samples. The other 42 samples contain discrepant predicted labels, where the LLMs revert to behaviors ingrained by previous training.

To simulate real-life deployment, we solicit 14 crowdworkers from diverse demographic backgrounds on the platform Amazon Mechanical Turks (MTurks) to annotate their preference on the explanations offered by the LLMs in this study . Annotators are encouraged to evaluate the pairs of explanation for each post based on the criteria of *Clarity, Reasoning* and *Alignment* as illustrated with the template in Figure 3. To avoid position bias, we present the pairs of explanations in randomized orders. All annotators are compensated fairly.

Label	T5	Llama	Row Total
Normal	25 (25.5%)	73 (74.5%)	98
Offensive	63 (49.6%)	64 (50.4%)	127
Hate	54 (46.2%)	63 (53.8%)	117

Table 2: Results of preference for explanations by MTurk annotators based on majority voting for 342 *HateXplain* samples.

Annotation Results Table 2 shows the annotation results, aggregated per sample by majority voting. Across the 3 labels in *HateXplain*, annotators show no significant preference for explanations by either model on *Offensive* and *Hate*. On the other hand, *Llama*'s explanations for *Normal* posts are overwhelmingly preferred.

6.2 Cross-model Refinement

6.2.1 Cross-model Training Methodology

Since *Llama* produces outputs preferred by human annotators, a natural choice is to select this model for ultimate deployment. We are nevertheless interested in examining the potential to steer one model's explanation style to align with another model's while maintaining existing classification performance.

Using the prompting technique described in section 6.1, we use the SFT+DPO $K = 128$ variants of both *Llama* and *T5* LLMs to generate explanations conditioned on the gold labels of the held-out 128-shot portion of the training data, which neither of these models has seen. We then perform cross-model SFT and DPO finetune using this dataset. More specifically, we further finetune the existing SFT+DPO $K = 128$ *T5* variant using explanations produced by its *Llama* counterpart, and vice versa. This step consists of both separate SFT and then DPO self-augmentation as in section 5.1.

6.2.2 Cross-model Classification Results

Figure 2 shows the macro-F1 scores of both the original single-model SFT+DPO and cross-model finetuned variants on the test datasets.

Cross-model training enhances T5 while impairing Llama. Applying SFT training using expla-

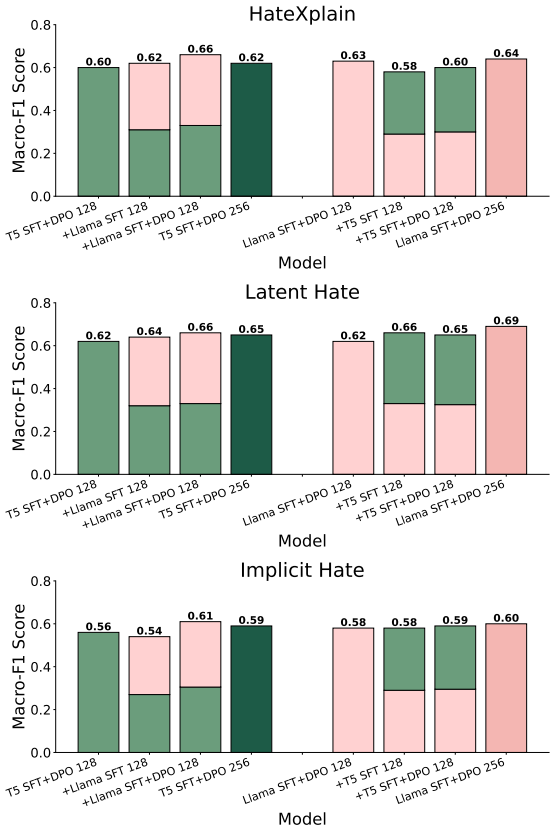


Figure 2: Macro-F1 scores on test portion of the 3 test sets for *T5* and *Llama* cross-model refinement regimen. In each figure: the first and last bars are scores reported with only *T5* model at $K = 128$ and $K = 256$ (directly from Figure 1); the middle bars are results after further finetuned using data from the complementary $K = 128$ shots of counterpart model. Split-color bars: bottom color indicates the original model; top color indicates the counterpart model for additional cross-training.

nations from *Llama* improves performance over the baseline *T5* SFT+DPO on *HateXplain* (F1 scores 0.62 from 0.60) and *Latent Hate* (0.64 from 0.62), with a slight reduction on *Implicit Hate* (0.54 from 0.56). On the other hand, applying SFT+DPO augmentation using *Llama*'s explanations from the complementary K=128 shot data allows T5 to **exceed** single-model performance on the full K=256 set across the board. The new maximum macro-F1 scores for *T5* are 0.66 on *HateXplain* (up from 0.62 of single-model), 0.66 on *Latent Hate* (up from 0.65), and 0.61 on *Implicit Hate* (up from 0.59). Notably, these scores are superior to even *Llama*'s single-model performance at $K = 256$ on *HateXplain* and *Implicit Hate*.

In contrast, SFT cross-training with *T5*'s output reduces *LLama*'s F1 scores to 0.58 from 0.63 on *HateXplain* while improving to 0.66 from 0.62 on *Latent Hate* and plateauing at 0.58 on *Implicit Hate*. Additional SFT+DPO cross-model augmentation slightly boosts performance on *HateXplain* and *Implicit Hate* while slightly impairing on *Latent Hate*. None of the cross-model *Llama* variants matches the single-models' F1 scores at $K = 256$.

6.2.3 Cross-model Style Analysis

We additionally train a style classifier based on the BERT architecture using explanations generated individually by *T5* and *Llama* of the SFT+DPO variant at $K = 128$ (Devlin et al., 2019). A classification head is added to BERT, which leverages the *[CLS]* token from each explanation's embedding to predict whether it was produced by *T5* or *LLaMA*. Technical details are reported in Appendix F.3.

From Figure 11, most explanations from the individually trained models exhibit their corresponding style, while cross-model SFT is sufficient to substantially align the outputs with the alternative's style. With the exception of *T5* on *HateXplain*, the application of self-augmenting DPO using cross-model data generally amplifies the stylistic presence of the base model to a small extent, suggesting that this mechanism induces a mild converging effect on the reasoning style.

Cross-model explanations may improve weak categories in weaker models. We compare *Precision*, *Recall*, and *F1-score* of base and cross-trained models at $K = 256$ (Figure 10) to assess cross-model refinement. Improvements often occur in categories where the paired model excels, indicating transferable reasoning patterns, while declines appear when the counterpart is weak. This asymmetry underscores our recommendation in line 19 of Algorithm 1 to validate and select the final model using a robust performance metric.

7 Analysis of Explanations

We perform Natural Language Inference (NLI) evaluation to ascertain whether the LLMs' explanations remain internally consistent with their corresponding labels and definitions throughout the augmentation pipeline, and then anatomize the stylistic changes between cross-model refinement.

7.1 Analysis of Consistency

Explanation-Label Consistency We prompt GPT-4o-mini with the template shown in Figure 8 to judge whether the offered explanations is logically consistent with the predicted label for each

Table 3: NLI-based consistency between explanations, predicted labels, and label definitions across datasets. Both DPO and cross-refined (XMOD) models show high entailment rates (>96%) with marginal percentages of contradictions (Contra.), neutral or undefined (Undef.) edge cases. Categories with all 0 values are omitted.

			Label Consistency (%)			Definition Consistency (%)		
Dataset	**Model**	**Train**	**Entail**	**Contra.**	**Undef.**	**Entail**	**Contra.**	**Neutral**
HateXplain	T5	DPO	99.2	0.8	0.0	98.2	1.5	0.3
		XMOD	96.7	1.5	1.8	99.0	0.9	0.2
	Llama	DPO	97.5	1.2	1.3	97.2	2.6	0.2
		XMOD	96.4	3.6	0.0	96.7	3.2	0.1
Latent Hate	T5	DPO	98.8	1.2	0.0	97.6	2.4	0.0
		XMOD	97.4	1.9	0.7	97.6	2.3	0.1
	Llama	DPO	97.3	2.7	0.0	97.3	2.3	0.3
		XMOD	96.8	3.2	0.0	97.5	2.5	0.0
Implicit Hate	T5	DPO	99.6	0.4	0.0	99.0	1.0	0.0
		XMOD	98.3	1.4	0.4	97.5	2.0	0.5
	Llama	DPO	98.1	0.4	1.5	97.5	1.8	0.8
		XMOD	96.9	2.8	0.1	96.0	3.6	0.4

sample across 3 test sets via the 4 categories: Entail, Contradict, Neutral and Undefined (noise cases where either the predicted label or the explanation is missing). An author's independent annotation of 150 random samples from this pool under the same instruction achieves nearly unanimous agreement with GPT-4o-mini (Cohen's $\kappa = 0.92$).

In Table 3, Entail values exceeding 96% indicate that *explanations are largely consistent with their respective labels*. However, cross-model training induces a marginal 2-3% increase in Contradiction compared to the DPO augmentation.

Explanation-Definition Consistency Applying the same method to *only the Entail samples* from the previous analysis, using the prompt shown in Figure 9 (Cohen's $\kappa = 0.96$), we again observe consistently high entailment rates and a slight increase in contradiction under cross-model training that mirrors the aforementioned trend.

Overall, LLM-generated explanations remain predominantly consistent with both predicted labels and their definitions. Cross-model refinement can enhance classification but may introduce a slight risk of inconsistency, warranting periodic monitoring. Appendix E offers qualitative analysis and insights for further quality checks for deployment.

8 Discussion and Conclusion

Informed by empirical insights , we offer the following recommendations for practitioners.

Cross-model refinement can further improve performance *T5*'s superior F1-scores compared to individually trained models (Figure 2) suggest added value in leveraging competing models. Practitioners can train multiple LLMs using the SFT+DPO framework across data thresholds, then select a high-quality checkpoint (e.g., $K = 128$ from Section 5.2.1) for further finetuning with outputs from other models. Although further gains are not guaranteed, cross-model refinement represents a practical and resource-efficient direction for improving alignment and robustness.

Architectural diversity may offer benefits. The differing post-training behaviors of *T5* and *Llama*—despite being trained on the same seed explanations—may stem from their distinct architectures (encoder-decoder vs. encoder-only). We hypothesize based on prior work (Hao et al.), that the stronger model (*Llama*) enhances reasoning diversity that *T5* can leverage. While further study is needed for a definitive explanation, our findings suggest practitioners may benefit from using architecturally diverse models within this framework.

Human oversight is still essential. Beyond the initial annotation, human supervision is necessary to monitor the stylistic quality of explanations, especially in iterative-development settings where unexpected patterns may occur after multiple rounds of cross-model training.

We hope that SMARTER will enable practitioners to effectively and efficiently deploy explanation-based content moderation pipelines.

9 Limitations

We discuss several limitations in our research, along with their implications for future explorations.

Generalization beyond English Our work focuses exclusively on English corpora. In reality, social media exists for every language. As both English-based LLMs and research literature often attract much more attention, we hope this framework will be adapted for other languages.

Model selection While we perform extensive experiments, our selection of 2 models could be expanded. At the cost of more training resource, training more models with different architectures or generation offer extra options for performance comparison for mix-and-match cross-training. We encourage other researchers to experiment with more LLMs using our framework.

Limited scope of human validation Due to budget constraints, we are only able to human validation for 1 dataset. In practice, it is recommended to solicit a larger pool of annotators to ensure appropriate amount of human oversight.

Risk of bias in automated explanation While our framework aims to improve performance and versatility in content moderation, it does not alleviate the fundamental sensitive nature of this task. Content moderation could be misappropriated to suppress free speech or harm marginalized groups (Dias Oliva, 2020; Kozyreva et al., 2023). Other works have noted that LLMs exhibit bias behaviors in certain applications (Nguyen et al., 2023; Gu et al., 2025). We urge researchers and practitioners to maintain vigilance when integrating our framework with human supervision to ensure ethical standards (Lai et al., 2022; Cao et al., 2024).

10 Ethical Considerations

The authors are not aware of any ethical problems in the development of this work. This research uses publicly available dataset. We also further synonymize the content to the best of our ability to provide extra caution for privacy. We acknowledge that our framework and associated techniques could be abused for harmful purposes.

References

Anjum and Rahul Katarya. 2024. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, 23(1):577–608.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242.

Stephanie Alice Baker, Matthew Wade, and Michael James Walsh. 2020. <? covid19?> the challenges of responding to misinformation during a pandemic: content moderation and the limitations of the concept of harm. *Media International Australia*, 177(1):103–107.

Dmitriy Bespalov, Sourav Bhabesh, Yi Xiang, Liutong Zhou, and Yanjun Qi. 2023. Towards building a robust toxicity predictor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 581–598.

Yang Cao, Lovely-Frances Domingo, Sarah Gilbert, Michelle Mazurek, Katie Shilton, and Hal Daumé Iii. 2024. Toxicity detection is not all you need: Measuring the gaps to supporting volunteer content moderators through a user-centric method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3567–3587.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.

Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and violent behavior*, 58:101608.

Peter Christen, David J Hand, and Nishadi Kirielle. 2023. A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):1–24.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2024. Is explanation all you need? an expert survey on llm-generated explanations for abusive language detection. In *Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*.

Thiago Dias Oliva. 2020. Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review*, 20(4):607–640.

Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Katharine Gelber. 2021. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*.

Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *arXiv preprint arXiv:2503.06778*.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hyesun Hwang and Kee-Ok Kim. 2015. Social media as a tool for social movements: The effect of social media use and social capital on intention to participate in social movements. *International Journal of Consumer Studies*, 39(5):478–488.

Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddia, Nabil Arhab, et al. 2024. A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms. *arXiv preprint arXiv:2404.00303*.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708.

Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120.

Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *arXiv preprint arXiv:2403.08035*.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of llms in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Huy Nghiem and Hal Daumé. 2024. HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5938–5956, Miami, Florida, USA. Association for Computational Linguistics.

Huy Nghiem, Umang Gupta, and Fred Morstatter. 2024. "define your terms": Enhancing efficient offensive speech classification with definition. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1309.

Thu T Nguyen, Junaid S Merchant, Xiaohe Yue, Heran Mane, Hanxue Wei, Dina Huang, Krishik N Gowda, Katrina Makres, Crystal Najib, Huy T Nghiem, et al. 2024. A decade of tweets: Visualizing racial sentiments towards minoritized groups in the united states between 2011 and 2021. *Epidemiology*, 35(1):51–59.

Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. Towards conceptualization of "fair explanation": Disparate impacts of anti-asian hate speech explanations on content moderators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9696–9717.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Melissa Kazemi Rad, Huy Nghiem, Andy Luo, Sahil Wadhwa, Mohammad Sorower, and Stephen Rawls. 2025. Refining input guardrails: Enhancing llm-as-a-judge efficiency through chain-of-thought fine-tuning and alignment. *arXiv preprint arXiv:2501.13080*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fbert: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(4).

Pabba Sumanth, Syed Samiuddin, K Jamal, Srikanth Domakonda, and Pathi Shivani. 2022. Toxic speech classification using machine learning algorithms. In *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, pages 257–263. IEEE.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Steven Windisch, Susann Wiedlitzka, Ajima Olaghere, and Elizabeth Jenaway. 2022. Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell systematic reviews*, 18(2):e1243.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, pages 500–507.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. 2025. Self-adapting language models. *arXiv preprint arXiv:2506.10943*.

Appendix

A Amazon Turks Mechanical Annotation

We obtained IRB approval before soliciting annotation from the crowd workers. 14 workers were recruited to annotate the 342 explanations introduced in subsection 6.1. Samples are assigned at random by the AMT platform to each annotator. Due to budget constraint, we opt to collect annotation for *HateXplain* to preserve statistical power instead of spreading across all 3 datasets. The breakdown of the annotators' demographic is below:

- ◇ **Gender**: Female (10), Male (3), Non-binary (1)
- ◇ **Age**: 18-29 (2), 30-39 (5), 40-49 (5), 50+ (2)
- ◇ **Race/Ethnicity**: Asian (3), Black (2), Hispanic/Latino (1), White (8)
- ◇ **Country of residence**: United States (14)
- ◇ **Education level**: 2-year college or equivalent (4), 4-year college or equivalent (6), High school or equivalent (3), Master degree or above (1)

B Data Pre-processing

The datasets used in this work are open-source. We peruse them as research artifacts accordingly to their license. To further mitigate risks of privacy, we anonymize posts by replacing user handles with the string *<user>* and remove URLs if they appear in the context of the posts.

We obtain the seed explanations from the repository associated with Nghiem and Daumé (2024) and merge them with the correct identifier keys.

C Obtaining Explanations for In-domain Data

The prompt template in Figure 7 is used to collect explanation conditioned on a single label. Note that more labels (and their definitions) may also be added to this section as desired.

D Stylistic Analysis

In Figure 11, we show the distribution of *Llama* vs *T5* style pre- and post cross-model refined as recognized by the BERT style classifier.

E Qualitative Analysis of Explanation

We examine the relatively small amount of non-Entail samples in section 7 to observe general trends in the models' inconsistency with respect to both the predicted label and definition.

First, *T5 models typically exhibit higher consistency in explanation compared to their Llama counterparts* as shown in Table 3. Even after cross-model refinement (XMOD), the percentage of Contradiction increases by at most 1% (*Implicit Hate*). While a conclusive explanation is out of scope, we hypothesize that *T5*'s encoder-decoder architecture supports higher robustness compared to *Llama*'s decoder-only.

Second, qualitative analysis reveals different patterns of inconsistency in the explanations offered by these 2 models. For *HateXplain* and *Implicit Hate*, *Llama*'s Contradict-tagged explanations tend to consistently mentions some negative and/or offensive sentiments in the post, yet ultimately yielding the conflicting label *Normal* (Figure 15). In contrast, *T5* Contradict explanations – much fewer in overall number – generally asserts a label at the beginning yet explicitly rejects this label at the end. For the more challenging *Implicit Hate*, both models' Contradict explanations typically contain "lost in context" inconsistency, where the justification is grounded in some other labels' definition rather than the stated one, possibly due to the nuanced complexity of these finer-grained categories.

Our NLI-based consistency check also emerges as a simple yet useful quality check for downstream deployment. In addition to classification metrics, moderators may opt to implement a secondary lightweight NLI-checker to alert when a threshold of inconsistency is reached.

F Technical Details

F.1 Hardware

All experiments are carried out on a maximum of 2 Nvidia RTX A6000 GPUs. We download the models from their Huggingface[3] checkpoints. Finetuning is implemented with LoRA techniques (Hu et al., 2021) via the TRL library [4]. The LoRA configurations for all settings are:

- ◇ Rank: 64
- ◇ *alpha*: 128
- ◇ Dropout: 0.05
- ◇ Target modules: q and v in projection layers

F.2 Inference Parameters

All models in our experiments use the following parameters during inference:

[3] https://huggingface.co/

[4] https://huggingface.co/docs/trl/en/index

Dataset	Total Size	Split Ratio	K val	K test	Platform	Label Space
HateXplain	20,148	60:20:20	50	400	G, T	Normal, Offensive, Hate
Latent Hate	19,112	60:20:20	50	400	T	Not Hate, Explicit Hate, Implicit Hate
Implicit Hate	4,153	50:20:30	50	150	T	White Grievance, Incitement to Violence, Inferiority Language, Irony, Stereotypes and Misinformation, Threatening and Intimidation

Table 4: Datasets for evaluating our framework. *Split Ratio* designates the proportion of the train:validation:test split. *K val* and *K test* are the number of data points per class to perform K-shot sampling from the Validation and Test set. For *Platform*, *G* denotes Gab, *T* for X (formerly Twitter).

Dataset	Model	Macro F1
HateXplain	Llama_OTS	0.52
	T5_OTS	0.59
Latent Hate	Llama_OTS	0.53
	T5_OTS	0.42
Implicit Hate	Llama_OTS	0.32
	T5_OTS	0.33

Table 5: Comparison of off-the-shelf (OTS) model performance across datasets. LLMs are prompted to produce answers without explanations in this setting as shown in Figure 5.

- temperature: 0.0
- max_tokens/max_new_tokens:
  - 512 for classification with explanation
  - 20 for classification without explanation

The library vLLM[5] is used coupled with the openAI `v1/chatcompletion` endpoint to enable fast inference in our experiments. The OpenAI model `GPT-5-chat-latest` endpoint was accessed on September 12, 2025. We also use the `Qwen2.5-32B-Instruct-AWQ` version on HuggingFace.

F.3 Preference Optimization

Direct Preference Optimization DPO is an alignment technique used to fine-tune large language models (LLMs) to better align with human preferences. Unlike traditional methods such as Reinforcement Learning from Human Feedback (RLHF), DPO simplifies the process by directly optimizing the policy model using a binary cross-entropy objective, thereby eliminating the need for a separate reward model and the complexities of reinforcement learning. This approach reparameterizes the reward model such that its optimal policy can be expressed in a closed form, transforming preference optimization into a classification problem (Rafailov et al., 2024).

Kahneman-Tversky Optimization KTO is another model alignment technique that aims to directly maximize the utility of a model's generations by applying principles from Kahneman-Tversky prospect theory (Kahneman and Tversky, 2013). KTO evaluates outputs based on their perceived gains or losses relative to a reference point, incorporating the concept of loss aversion. This method utilizes data that simply indicates whether an output is desirable or undesirable for a given input, which can be more readily available and less expensive to collect than the paired preference data required by DPO (Ethayarajh et al., 2024).

Implementation We primarily use the implementation of the `DPOTrainer` and `KTOTrainer` classes via the TRL library hosted on the HuggingFace platform. For the `DPOTrainer`, we set $\beta = 0.1$, and the loss to be default *sigmoid* loss. For the `KTOTrainer`, we similarly set the $\beta = 0.1$ and use the default `KTO loss` option.

F.4 Hyperparameter Tuning

For ModernBERT, we use the HuggingFace Trainer and train for 4 epochs with AdamW optimizer. We use the associated tokenizer and pad the inputs to maximum length (512) for each batch.

For Base SFT finetuning, we use 3 epochs and learning rate 3e-4. We present the chosen parameters for DPO/KTO alignment tuning in Table 6, Table 7, Table 8 below.

G Examples of Explanations

We provide some examples of the explanations generated during our cross-training phrase in Figure 12, Figure 13, and Figure 14.

[5] https://docs.vllm.ai/en/latest/

K	Technique	T5 Epochs	T5 LR	Llama Epochs	Llama LR
16	DPO	3	5e-05	3	1e-05
32	DPO	3	5e-05	4	1e-05
64	DPO	3	1e-05	3	1e-05
128	DPO	3	5e-05	3	1e-05
256	DPO	4	1e-04	3	1e-05
256	KTO	–	–	3	5e-07

(a) Hyperparameters for regular training.

K	Technique	T5 Epochs	T5 LR	Llama Epochs	Llama LR
256	DPO-K128	3	1e-05	4	1e-04
256	DPO-K192	3	1e-04	1	7e-05
256	DPO-N128	3	5e-05	5	5e-06
256	DPO-N192	3	5e-05	5	5e-06

(b) Hyperparameters for subsampling training.

Table 6: *HateXplain*.

K	Technique	T5 Epochs	T5 LR	Llama Epochs	Llama LR
16	DPO	3	5e-05	3	1e-06
32	DPO	3	1e-04	3	1e-06
64	DPO	3	5e-05	3	1e-06
128	DPO	4	5e-05	3	1e-04
256	DPO	3	1e-04	3	1e-04
256	KTO	–	–	3	5e-07

(a) Hyperparameters for regular training.

K	Technique	T5 Epochs	T5 LR	Llama Epochs	Llama LR
256	DPO-K128	4	1e-04	4	1e-04
256	DPO-K192	3	1e-04	3	1e-04
256	DPO-N128	4	1e-04	3	1e-04
256	DPO-N192	4	1e-04	5	1e-04

(b) Hyperparameters for subsampling training.

Table 7: *Latent Hate*.

K	Technique	T5 Epochs	T5 LR	Llama Epochs	Llama LR
16	DPO	3	5e-05	3	5e-06
32	DPO	3	5e-05	4	5e-05
64	DPO	3	1e-04	3	1e-06
128	DPO	1	7e-05	3	1e-04
256	DPO	1	5e-05	1	5e-05
256	KTO	–	–	3	5e-07

(a) Hyperparameters for regular training.

K	Technique	T5 Epochs	T5 LR	Llama Epochs	Llama LR
256	DPO-K128	1	7e-05	1	1e-05
256	DPO-K192	1	7e-05	1	5e-05
256	DPO-N128	1	7e-05	1	1e-05
256	DPO-N192	1	7e-05	1	5e-05

(b) Hyperparameters for subsampling training.

Table 8: *Implicit Hate*

Please read instruction carefully before attempting !

Instruction

Labels and Definitions

Instructions:

You are presented with a Post, a set of Labels and their Definitions, and two explanations EXP_A and EXP_B.

Your overall objective is to evaluate both explanations and choose the one that better provides a clear, logical, and convincing justification for the chosen label.

To achieve this, follow these steps:

- Review the entire set of labels and their definitions. Ensure that the label chosen in each explanation aligns with the content of the post.
- Evaluate both explanations based on the following criteria:
  - **Clarity:** How clearly is the reasoning communicated?
  - **Reasoning:** Is the reasoning behind the label choice logical and well-supported?
  - **Alignment:** Does the explanation adhere to the label's definition and apply it accurately to the post?
  - **Format:** The explanation should follow the format:
    EXPLANTION: [text] LABEL: [text]
- Finally, select the explanation that best fits the criteria above.

Labels and Definitions:

Hate: language which attacks, demeans, offends, threatens, or insults a group based on race, ethnic origin, religion, disability, gender, age, sexual orientation, or other traits. It is not the presence of certain words that makes the text hate speech, rather you should look for the context the word is used in the text.

Offensive: language that is rude, hurtful, derogatory, obscene or insulting, or to upset or embarrass people.

Normal: language that is neither Hate speech nor Offensive speech.

Post:

${post}

EXP_A:

${exp_a}

EXP_B:

${exp_b}

Select the better explanation:

○ EXP_A
○ EXP_B

Submit

Figure 3: Template to collect annotation for preference on explanations for Amazon Mechanical Turks crowdworkers.

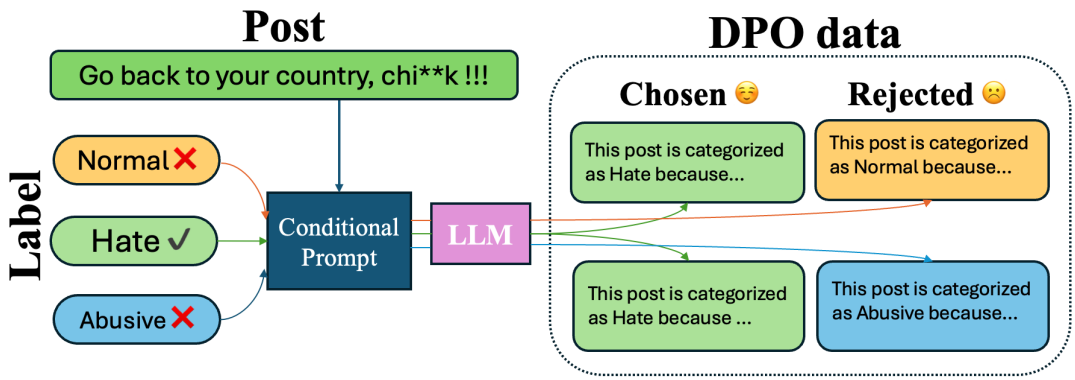


Figure 4: Self-augmenting pipeline: for each post, explanations are conditionally generated for the gold label and all incorrect labels using prompt template in Figure 6. For DPO, data consists of matching the explanation of the correct label with another incorrect label's.

```
### Instruction:
By considering their corresponding Definitions, label the following post with
only one of these categories: {categories}.

Provide your response in the following format:
LABEL: [text]

### Definitions:
{definitions}

### Post:
{post}

### Response: [model output]
```

Figure 5: Prompt template for classification tasks **without explanation generation**. The model is instructed to directly output a label based on the provided definitions.

```
### Instruction:
By considering their corresponding Definitions, label the following post with
only one of these categories: categories. Provide a brief paragraph to explain
step-by-step why the post should be classified with the provided Label based on
the given Definitions.
If this post targets a group or entity relevant to the definition of the specified
Label, explain who this target is/are and how that justifies the chosen Label.
Provide your message in the following format:
EXPLANATION: [text]
LABEL: [text]
### Definitions:
Hate: Language that targets [...]
Abusive: Language that is [...]
Normal: Does not contain [...]
### Post:
{post}
### Response: [model output]
```

Figure 6: Prompt template for classification tasks. The definition block should contain the set's full label space

```
### Instruction:
By considering the following Post, the Label and its Definition, provide a brief
paragraph to explain step-by-step why the post should be classified with the
provided Label based on the given Definition.
If this post targets a group or entity relevant to the definition of the specified
Label, explain who this target is/are and how that justifies the chosen Label.
Provide your message in the following format:
EXPLANATION: [text]
LABEL: [text]
### Definitions:
Hate: Language that targets [...]
### Post:
{post}
### Response:
```

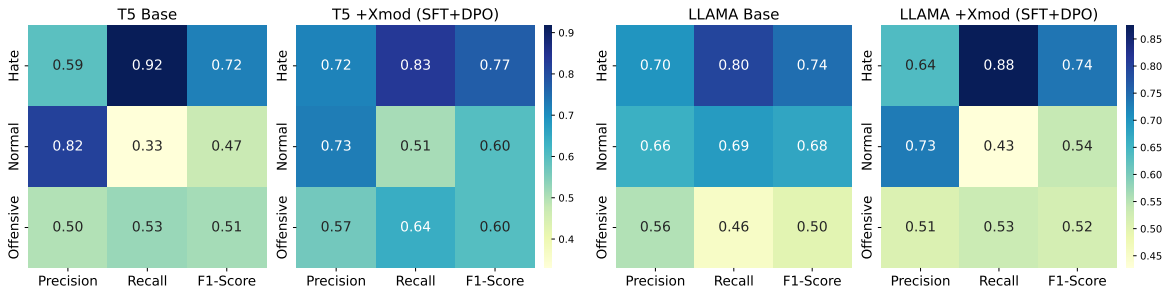
Figure 7: Prompt template to obtain explanation conditioned on a single label and its definition.

```
### Instruction:
You are given a model's prediction and its explanation for a social media post.
Your task is to judge whether the explanation is logically consistent with the
predicted label.

Possible outputs:
- ENTAIL: The explanation supports and is faithful to the predicted label.
- CONTRADICT: The explanation contradicts the predicted label.
- NEUTRAL: The explanation is unrelated, vague, or does not clearly support the
predicted label.
- UNDEFINED: Either the predicted label or the explanation is missing or empty.

Return only one word: ENTAIL, CONTRADICT, NEUTRAL, or UNDEFINED.

### Predicted Label:
{pred_label}

### Model Explanation:
{explanation}

### Response: [model output]
```

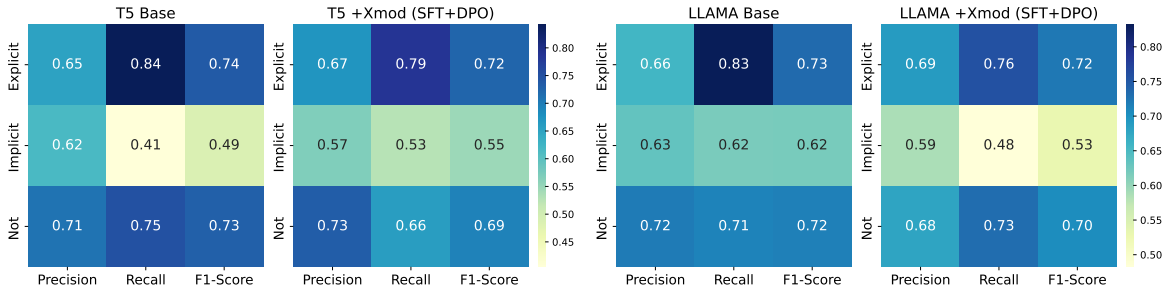
Figure 8: Prompt template for evaluating **explanation–label consistency**. The model is asked to judge whether a generated explanation is logically consistent with its predicted label.

```
### Instruction:
You are given a model's predicted label and its corresponding explanation.
Your task is to judge whether the explanation is logically consistent with the
definition of that label.

Possible outputs:
- ENTAIL: The explanation clearly aligns with and supports the definition.
- CONTRADICT: The explanation conflicts with the definition.
- NEUTRAL: The explanation is vague, unrelated, or does not clearly support the
definition.

Return only one word: ENTAIL, CONTRADICT, or NEUTRAL.

### Official Definition:
{definition}

### Model Explanation:
{explanation}

### Response: [model output]
```

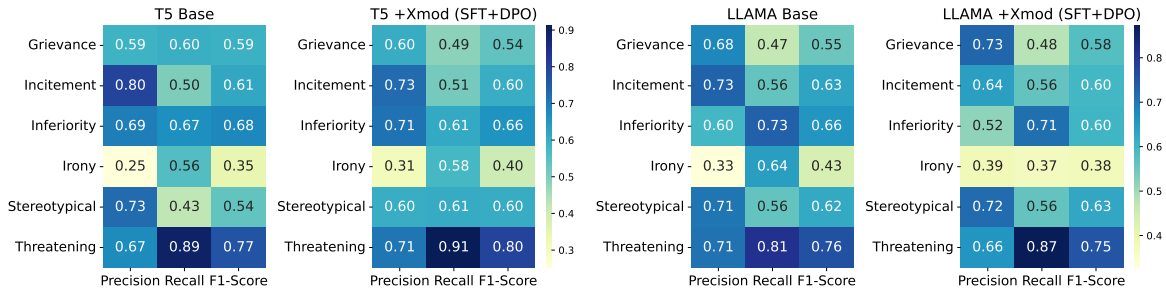
Figure 9: Prompt template for evaluating **explanation–definition consistency**. The model is asked to assess whether a generated explanation logically aligns with the formal definition of the predicted label.



(a) *HateXlain*

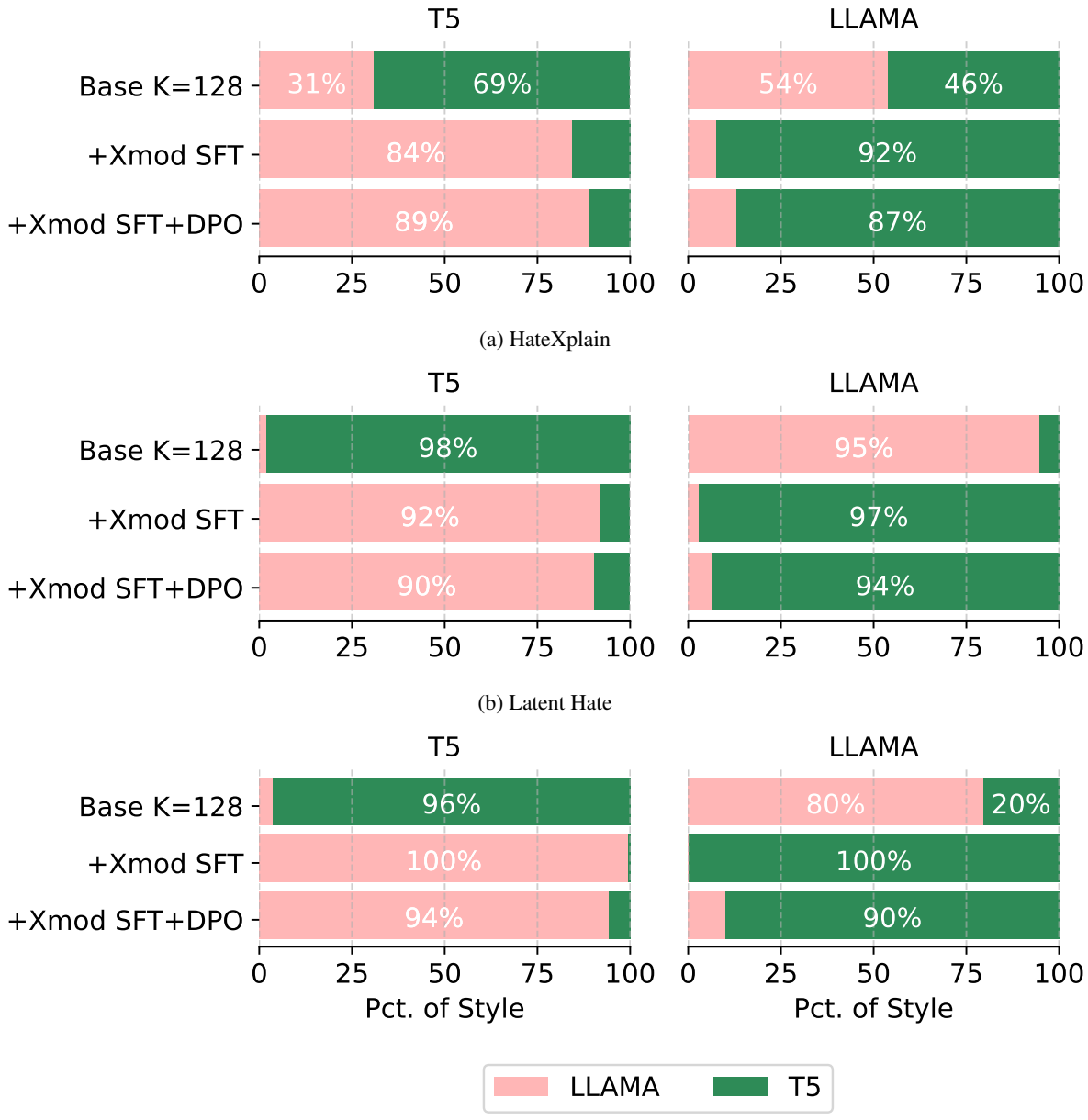


(b) *Latent Hate*



(c) *Implicit Hate*

Figure 10: Comparison of Precision, Recall, and Macro F1-scores for the *T5* and *Llama* model families under two training regimes: self-augmentation only (*Base*) and cross-model refinement (*+Xmod*) at $K = 256$. Models trained with cross-model refinement (right plot within each model family) generally show improvements in label categories where their counterparts already perform well, and conversely exhibit declines in areas corresponding to their weaknesses.



(a) HateXplain

(b) Latent Hate

(c) Implicit Hate

Figure 11: Percentage distribution of *Llama*- and *T5*-style explanations on test sets by BERT style classifiers, pre (*Base*) and post cross-model (*Xmod*) refined.

Dataset	Definition
HateXplain	◇ **Normal:** neither hate speech nor offensive. ◇ **Hate Speech:** language which attacks, demeans, offends, threatens, or insults a group based on race, ethnic origin, religion, disability, gender, age, sexual orientation, or other traits. it is not the presence of certain words that makes the text hate speech, rather you should look the context the word is used in the text. ◇ **Offensive Language:** usage of rude, hurtful, derogatory, obscene or insulting language to upset or embarrass people.
Latent Hate	◇ **Not Hate:** speech or actions that do not involve any form of hatred, prejudice, or discrimination toward individuals or groups based on their characteristics. ◇ **Explicit Hate:** openly expressed, direct forms of hatred and prejudice toward individuals or groups based on their characteristics. ◇ **Implicit Hate:** coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity.
Implicit Hate	◇ **White Grievance:** includes frustration over a minority groups perceived privilege and casting majority groups as the real victims of racism. This language is linked to extremist behavior and support for violence. ◇ **Inferiority Language:** implies one group or individual is inferior to another, and it can include dehumanization, denial of a person's humanity, and toxic language that compares the target with disease, insects, animals. Related to assaults on human dignity, dominance, and declarations of superiority of the in group. ◇ **Irony:** refers to the use of sarcasm , humor, and satire to attack or demean a protected class or individual. ◇ **Stereotypes And Misinformation:** associate a protected class with negative attributes such as crime, or terrorism. includes misinformation that feeds stereotypes and vice versa, like holocaust denial and other forms of historical negationism. ◇ **Threatening And Intimidation:** conveys a speaker's commitment to a target's pain, injury, damage, loss or violation of rights, threats related to implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation.

Table 9: Definitions of toxic label categories and definitions (collected from their original publications) across datasets.

POST:
what kind of brzen hapless *** uses an expression like "there's no big *** there"
LABEL: HATE

(a) A sample post in *HateXplain* with gold label. Offensive content partially redacted.

VARIANT: T5 DPO @ K=128
EXPLANATION: The post targets the group of homosexuals by using a derogatory term "***" to insult and demean them. This language is offensive as it is rude, hurtful, and insulting. Therefore, the post is labeled as Offensive.
LABEL: Offensive

VARIANT: T5 +XMOD SFT+DPO @ K= 256
EXPLANATION: The post targets the group of individuals who identify as "brzen" and use the derogatory term "***" The language used in the post attacks and demeans this group based on their sexual orientation, which falls under the definition of hate speech.
LABEL: Hate

VARIANT: LLAMA DPO @ K=128
EXPLANATION: The post targets the homosexual community by using a derogatory slur "***" to demean and insult them. This language is offensive as it is rude, hurtful, and insulting towards individuals based on their sexual orientation. Therefore, the post is labeled as Offensive.
LABEL: Offensive

VARIANT: LLAMA +XMOD SFT+DPO @ K=256
EXPLANATION: The post targets the homosexual community by using a derogatory term "***" to insult and demean them. This language attacks and demeans a specific group based on their sexual orientation, which falls under the definition of hate speech. Therefore, the post is labeled as Hate.
LABEL: Hate

Figure 12: Examples of a post in *HateXplain* with gold label, along with the explanations of the *T5* and *Llama* self-augmented variants at K=128, and the cross-trained model using all K=256 in total. We observe that models trained with more data (+Xmod) gives the correct classification.

POST:
it means people aren't buying the lie! diversity k**s now prove me wrong
LABEL: IMPLICIT HATE

(a) A sample post in *Latent Hate* with gold label. Offensive content partially redacted.

VARIANT: T5 DPO @ K=128
EXPLANATION: This post does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. It is a statement expressing skepticism towards the belief that diversity ***, but it does not disparage or express any form of hate towards any specific person or group. Therefore, it can be classified as "Not Hate."
LABEL: Not Hate

VARIANT: T5 +XMOD SFT+DPO @ K= 256
EXPLANATION: This post does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. The post simply states that people are not buying the lie that diversity ***, and asks for proof to prove the opposite. There is no indication of any negative or discriminatory language towards any specific group or entity. Therefore, this post should be classified as "Not Hate."
LABEL: Not Hate

VARIANT: LLAMA DPO @ K=128
EXPLANATION: The post uses coded or indirect language to disparage a person or group based on their race. Overall, the use of the phrase "diversity ***" and the challenge to "prove me wrong" indicate a negative view of diversity, suggesting a belief that diversity is harmful. This coded language indirectly disparages people of color, making it an example of implicit hate.
LABEL: Implicit Hate

VARIANT: LLAMA +XMOD SFT+DPO @ K=256
EXPLANATION: This post should be classified as "Not Hate" because it does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. The post simply expresses a belief that diversity is a lie and challenges the reader to prove the speaker wrong.
LABEL: Not Hate

Figure 13: Examples of a post in *Latent Hate* with gold label, along with the explanations of the *T5* and *Llama* self-augmented variants at K=128, and the cross-trained model using all K=256 in total. In this example, we observe that *T5* absorbs the verbosity of *Llama* after cross-training, yet still gets the wrong label. On the other hand, *Llama*'s explanation is more terse after cross-training, and also classifies the post incorrectly.

POST: hi can i get a couple of grammes of your *** power ?

LABEL: INCITEMENT

(a) A sample post in *Implicit Hate* with gold label. Offensive content partially redacted.

VARIANT: T5 DPO @ K=128
EXPLANATION: The post targets minorities by expressing frustration over their perceived privilege and casting the majority group (*** people) as the real victims of racism. The post uses language that elevates *** power and unity, which is a known hate group ideology. The post also includes a sarcastic and ironic tone by asking for a "couple of grammes of your *** power," implying that minorities are inferior and need to be subjugated. This aligns with the definition of incitement, as it flaunts in-group unity and power while also elevating known hate groups and ideologies.
LABEL: Incitement

VARIANT: T5 +XMOD SFT+DPO @ K= 256
EXPLANATION: The post flaunts in-group unity and power by flaunting the idea of *** power. By using the term "*** power," the post elevates the idea of *** power and elevates the hate group ideology of *** supremacy. This aligns with the definition of incitement, which includes flaunting in-group unity and power or elevating known hate groups and ideologies.
LABEL: Incitement

VARIANT: LLAMA DPO @ K=128
EXPLANATION: This post targets *** people and their perceived desire for power and authority. By specifically asking for a couple of grammes of *** power, the post elevates and promotes the idea of *** supremacy, which is a hate group and ideology that advocates for the superiority of the *** race. This incitement of in-group unity and power aligns with the definition of incitement, as it promotes and supports a hate group and their extremist behavior.
LABEL: Incitement

VARIANT: LLAMA +XMOD SFT+DPO @ K=256
EXPLANATION: The post directly flaunts in-group unity and power by referencing *** power, which is a known hate group and ideology. The post specifically asks for a couple of grammes of *** power, which further promotes the idea of *** supremacy. This aligns with the definition of incitement, as it flaunts in-group unity and power and elevates known hate groups and ideologies.
LABEL: Incitement

Figure 14: Examples of a post in *Implicit Hate* with gold label, along with the explanations of the *T5* and *Llama* self-augmented variants at K=128, and the cross-trained model using all K=256 in total. In this example, we also observe the cross-pollination of style in terms of verbosity, albeit to a lesser degree, but all models classify correctly.

LLAMA - HATEXPLAIN

EXPLANATION: The post targets the group of homosexuals by using the term "queer" in a derogatory manner. The language used attacks and demeans this group based on their sexual orientation, which falls under the definition of hate speech. Therefore, the post is labeled as "Normal."
LABEL: Normal

(a) Llama: The final label 'Normal' does not follow from the prior justification.

T5 - LATENT HATE

EXPLANATION: This post should be classified as "Not Hate" because it does not involve any form of hatred, prejudice, or discrimination towards individuals or groups based on their characteristics. The post simply mentions Alex Linder and white nationalism 1.0, without expressing any negative or discriminatory views towards any specific group or individual. Therefore, it does not meet the criteria for being classified as "Not Hate."
LABEL: Not Hate

(b) T5: The explanation starts with an asserted label yet later refutes this category.

LLAMA - IMPLICIT HATE

EXPLANATION: The post uses irony to attack and demean the idea of a wall, sarcastically suggesting that it should also prevent merpeople from the ocean. The use of humor and satire is used to mock the idea of a wall being a barrier to prevent all forms of transportation, including boats, tunnels, and airplanes, as well as the ocean.
LABEL: Threatening

(c) Llama: The explanation discusses factors directly relevant to 'Irony', yet the predicted label 'Threatening' is unrelated.

Figure 15: Examples of explanations that are judged to contains some forms of inconsistency with respect to the predicted label and/or its definition in our NLI-based analysis.